DOCUMENT RESUME

ED 206 737 TM 810 688

AUTHOR Wiley, David E.
TITLE The Vicious and the Virtuous: ETS and College Admissions.
INSTITUTION Northwestern Univ., Evanston, Ill.
SPONS AGENCY National Academy of Education, Washington, D.C.; National Inst. of Education (ED), Washington, D.C.
PUB DATE Jan 81
GRANT NIE-G-78-0155
NOTE 47p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Book Reviews; *College Entrance Examinations; Test Format; Testing; *Testing Problems; *Test Validity
IDENTIFIERS *Educational Testing Service; Scholastic Aptitude Test; Testing Industry

ABSTRACT

The Ralph Nader report on the Educational Testing Service (ETS), entitled The Reign of ETS, the Corporation That Makes Up Minds, explicitly and implicitly raises serious issues concerning the testing enterprise. Major themes include the role of testing in the educational selection system, the validity of existing tests, and the corporate power of ETS. This paper explores and evaluates charges from the perspectives of social and educational policy and of psychometric theory and practice. The currently polarized controversy over testing is reflected: advocates of testing believing in the validity and social utility of both the science and technology of mental measurement, and critics proclaiming that technical inadequacies and social harms make the enterprise inherently vicious and invidiously misleading. The paper concludes that the tests are central to the sorting of individuals into successive types of education. It advocates a more open procedure for specifying the content of selection tests, and concludes that current psychometric conceptions are inadequate for a meaningful assessment of the validity of tests. (Author/GK)

THE VICIOUS AND THE VIRTUOUS: ETS AND COLLEGE ADMISSIONS

David E. Wiley
Northwestern University

January, 1981

The Reign of ETS: The Corporation that makes up minds, by Allan Nairn and Associates. Washington, D.C.: Ralph Nader, 1980. xv + 554 pp. $50.00.

This work was supported by the National Academy of Education and by the National Institute of Education (Grant No. NIE-G-78-0155). The opinions expressed in this publication do not necessarily reflect the positions, policies, or endorsements of the supporting organizations.

ABSTRACT

The Ralph Nader Report on the Educational Testing Service, entitled The Reign of ETS, the Corporation that makes up minds, is authorized by Allan Nairn and his associates. Both, explicitly and implicitly, it raises serious issues concerning the testing enterprise. Major themes include the role of testing in the educational selection system, the validity of existing tests -- in terms of appropriateness of tested skills as well as the accuracy of the measurements -- and the corporate power of ETS. This paper explores and evaluates charges from the perspectives of social and educational policy and of psychometric theory and practice.

The report reflects the currently polarized controversy over testing: Advocates of testing strongly believing in the validity and social utility of both the science and technology of mental measurement and critics loudly proclaiming that technical inadequacies and social harms make the enterprise inherently vicious and invidiously misleading.

The paper concludes that the tests are central to the sorting of individuals into successive types of education. This centrality makes their content and use key foci for political debate. The paper then discusses aptitude and achievement as educational concepts and interprets the content of the Scholastic Aptitude Test in these terms. It advocates a more open procedure for specifying the content of selection tests. The paper also concludes that current psychometric conceptions are severely inadequate for a meaningful assessment of the validity of tests, and that, in particular, serious questions concerning potentially distorting influences of multiple choice formats are inadequately addressed within existing frameworks.

The amount of time American children and youth devote to being examined with pre-prepared testing materials is unknown. But, consensually, it is immense. Currently, there is surely no young person in the United States who, by high school age, has not participated in a "standardized" testing. The extent of this activity and the perceived importance of the information provided by it for educating pupils have, in recent years, provoked serious discussion of the worth of the endeavor for both the individuals involved and the society as a whole.

The enterprise has vigorous advocates and staunch defenders. These include scholars and educational practitioners as well as those who devote their professional lives to the development and dissemination of testing materials and procedures. The strong claims made by these individuals include:

(1) That there has been a distinguished history of scientific and technical development which has produced a technology of mental measurement allowing the valid assessment of important mental traits. And these traits range from the most fundamental--i.e., intelligence--to the most specific and immediate--e.g., knowledge or skills resulting from particular school lessons.

(2) The applications of this technology in instruction, evaluation, and selection has resulted in great

> benefits to individuals and to the society, and specifically, it has made the process of schooling more virtuous because it has enhanced the accuracy of information used to decide which instruction would be most effective for particular pupils, which methods and programs are generally superior or how they might be improved, and which pupils are sufficiently meritorious that they should have opportunities for continued and higher-level education. And with regard to the latter measuring devices, the claim is that the transition from "subjective" to "objective" procedures actually resulted in substantially greater access of disadvantaged and minority pupils to further schooling because the "subjective" decision processes used formerly were biased in favor of advantaged pupils from middle-class backgrounds.

On the other side, however, there are critics and attackers. As with most who are critical of what has become an entrenched establishment, they are shrill and occasionally--some would say often--impolite. They tend to be political rather than academic in their rhetoric, because they know that polite academic discussions draw the attention of polite academics, but seldom that of others--at least in the short run. And no one can

produce social change without grasping the attention of those who must originate and manage that change.

The core content of their message is serious and in direct opposition to the claims of the advocates:

(1) The tests are invalid. They purport to measure important mental characteristics and they, in fact, measure trivial ones. They are individiously biased against those coming from disadvantaged backgrounds, reflecting their capacities even more inaccurately than those of individuals with middle class origins.

(2) The wide-spread use of these tests for instruction, evaluation, and selection has been harmful to both the individuals tested and to the society. Specifically, it has made the process of schooling more vicious, because it has biased the educational decisions made about pupils. In particular, some pupils are wrongly labeled as mentally handicapped and are invalidly excluded from instruction potentially beneficial to them. Many are inappropriately tracked and labeled as mentally inferior so continuously and so publicly that they internalize the label and are permanently emotionally and motivationally scarred. Finally, numerous pupils are illegitimately barred from subsequent higher-level education. And with

> regard to the measuring devices used for these latter "selection" decisions, the test constructors have chosen narrowly defined, trivial, and culturally selective mental content, further distorted in the measurement process by multiple-choice formats, resulting in a claimed scale of uniform, universal mental merit, which even when predictively "validated" against a seriously inadequate criterion, exhibits only the weakest of relations. These devices, therefore, bias the selection process against those whose talents are underrepresented in, or distorted by the tests. In particular, economically disadvantaged and racial and cultural minorities are most strongly injured, and their already deprived standing is reinforced by the selection processes incorporating these devices.

The criticism of tests and the controversy over the impact of testing is not an isolated phenomenon. Few areas of societal activity are currently free of conflict. The concerns that were originally focused on profit-making corporations in terms of product safety and the environmental impact of industrial processes caused so much to be revealed that shocked and mobilized public opinion that now no societal institution is presumed virtuous. From the professions to the presidency all is subject to scrutiny and critique.

Surely one primal impetus to all of this has been Ralph Nader. His main concern and that of those around him has been the consumer and the corporation. Corporations have a responsibility to those who purchase their wares as well as to their stockholders and Nader has taken as his task insuring that they meet these responsibilities. Unsafe at Any Speed served as a critical model for what can be accomplished and how to accomplish it.

Following that model, the book under review focuses on tests and testing, their impact on their consumers--defined as those who take them--and the corporation that dominates the field--the Educational Testing Service.

As Nader indicates in the preface, the notion behind the book was his. It was a ripe area and time for articulating consumer concerns: there was and is intense anxiety on the part of students about ETS selection tests; they play a central role in access to the educational experiences required of individuals who would occupy key leadership positions in the society; there is and has been respectable legal and academic criticism of these products and their use; the "industry" is dominated--technically, if not financially--by a single corporation.

After some preliminary studies, the project was initiated in 1974 under the auspices of the New Jersey Public Interest Research

Group. Allan Nairn coordinated a small group which collected information from a reluctant ETS, and waged an active media campaign to stimulate corporate responsiveness and alternate information sources. After a long gestation period, the book under review emerged.

The book begins (Chapter 1) with a set of quotations from Henry Chauncy and other founders of ETS, exposing their grand dreams for the influence of the organization, with heavy emphasis on how a great expansion of scientific testing could better allocate and select talented individuals and minimize unrealistic aspirations. This is followed by a series of case descriptions of how individuals (consumers) have been grievously injured by the selection tests and procedures. Chapter 2 documents the size and power of the corporation focusing on the opulence of the facilities and the influence of the corporation and its board of executives. It follows this with description of all ETS testing programs organized by age or educational level of person tested--from tests used for preschoolers to those designed for graduate/professional school selection.

Well over one quarter of the book is contained in Chapter 3 which exposes the heart of the criticism of tests and testing. It focuses on test validity beginning with the idea of predictive validation and discussing appropriate quantitative

indices and concluding that the predictive quality of ETS section tests is so low as to be trivial. This is followed by a critique of the validity of the predicted criterion itself--first year grades; making the point that more obviously valid criteria--later grades, subsequent post-school success, etc., are even more poorly forecast. Subsequently it presents a more substantive attack on validity; discussing test anxiety, risk taking strategy, including guessing and coaching, among other topics. After this, attention is directed toward economic and ethnic differences in test scores with extensive discussion of various components of test bias: experientially or culturally biased item content; omission from testing of individual characteristics which are more predictive of later life success and which do not exhibit such great group differentials; the conclusions of the new literature on test bias in selection, focusing on selection bias against low scoring groups which result from low test validity; and generally making the point that the admissions testing system reinforces social inequality over generations by denying low scoring groups the opportunity to improve their educational preparation. Finally, the chapter continues with discussion of secrecy and truth in testing, the process and the small amounts of resources devoted to construction of multiple choice instruments, concluding that actual multiple choice test development costs for ETS admissions programs vary from 2 percent to 9 percent of the amounts charged candidates. The ultimate comments

of the chapter then focus on how the test equating and scaling system used for the tests "mystifies" the meaning of the scores and how small performance differences--in terms of number of items correctly answered--can make for differences in scale scores which have important impacts on candidates' chances for admission.

Chapter 4 is an historical review of the testing movement and its main intent is to link the work of Carl Brigham--the originator of the modern multiple choice SAT--to the early metal testing movement and its contribution to genetic arguments of the mental inferiority of racial groups as they entered political argumentation over immigration policy. Chapter 5 is a review of the relation between social class--income, occupation of parents--and score levels on aptitude tests--primarily the SAT. Claiming the relation is stronger than the predictive validity relation, the discussion focuses on how the use of the test for admission systematically excludes working class youth from the educational experiences necessary for life success and societal influence. Chapter 6 concentrates on the Law School Aptitude Test, emphasizing the key leadership role that lawyers play in the society and the fact that the LSAT is required of all who would attend U.S. Law Schools. Chapter 7 outlines the peculiar nature of the contract entered into by the consumer and ETS, emphasizing the almost total lack of rights retained by the person tested.

The second most extensive section of the book (Chapter 8) is devoted to an analysis of ETS as a corporate entity. It puts forth the view that ETS acts precisely like a profit-making corporation, but that it has no stockholders to which it is accountable and thus is not subject to control. It discusses the relations which ETS has to other entities--the College Board, client organizations --and strongly emphasizes the way in which ETS' prestige and institutional linkages reinforce and increase its market dominance, both domestic and international. The discussion then turns to analysis of financial records to show that ETS does in fact, make a "profit" and how it uses it. Finally, it analyzes the payroll and employment records to exhibit the pay differentials of professional and non-professional staff, compares compensation with universities and other organizations and concludes that managerial personnel are overpaid and that those in clerical and other non-professional positions--where minority workers are concentrated--are grossly underpaid. This section also emphasizes the lack of success that organizational features intended to increase minority participation and influence at the corporate policy level have had.

The concluding Chaper (8) sums up and raises five "policy" questions for public discussion. These revolve around: (a) the desirability of an admissions information system focused on multiple choice tests and concentrated in a single organization, (b) the admissions policies that the tests serve and their intended and unintended con-

sequences for individuals and the society, (c) whether the direct costs should be borne by the institutions rather than the individuals, (d) the meaning of the score scale and the issue of whether alternate criteria should be scaled at all, (e) the possibility of reform given the centrality of ETS to the whole testing enterprise. This chapter, however, places an optimistic vision over the issues, beginning its discussion with a salute to The New York truth-in-testing legislation.

The book has three dominant themes

--the importance of the educational selection system to the individuals involved and to the society and the accusation that the existing system, centrally involving current admissions tests, is both psychologically and socially harmful;

--the invalidty of the admissions tests, in terms of what they measure, and how they measure it and the resulting inequities for the disadvantaged and for ethnic minorities;

--the corporate power and influence of ETS as it increasingly dominates the private and public domestic testing markets and as it expands its influence abroad. Concomitant with this charge is the difficulty of society's controlling an industry-dominant not-for-profit corporation when it chooses to act like a profit-making

entity--accumulating and investing earnings--rather than like a charitable or educational institution.

The remainder of this review will focus primarily on two of these themes--the tests themselves and the system of educational and social differentiation within which they are embedded.

The tests

For the testing enterprise, the report's most serious charge has directly to do neither with the societal selection process within which tests are embedded nor with the corporate power of ETS. Neither of these are central to either the intellectual rationale for or the activities which constitute test creation and use. The core charge, for those in testing, concerns the validity of the tests themselves.

Validity, for Nairn and his associates, has (implicitly) a wide scope. It incorporates questions concerning the societal worth, legitimacy, and centrality of the human characteristics assessed by the tests. It encompasses issues relating to the distortions induced by the testing procedures, assumptions and methods of scoring and scaling; and modes of presentation which cause the results of the assessments to be discrepant from the characteristics assessed. And it incorporates the issue of typical differences in these discrepancies between social groups, i.e., what the psychometric community has come to call "test bias."

These three areas of concern about tests are, however, hardly differentiated in the text of the report. The discussion there raises many sub-issues of a critical nature but organizes them in such a diffuse fashion that the more general criticisms are hard to fathom and difficult to distil, let alone accept or rebut. The reasons for this diffuseness are several: the polemical, rather than analytical, intent of the report; the lack of conceptual grounding of the argument--which makes it difficult to bridge the gap between social concern and technical analysis; and inherent vaugeness in the traditional notion of validity, as it is currently interpreted in the psychometric community.

The tests: their content

The social needs to be met by the processes controlling access to higher education are many. One wishes to select individuals who will "do well" and thus succeed in the enterprise without consuming more of institutional and societal resources than we can afford to allocate. One wishes to "reward" hard work and, especially, academic accomplishment at earlier stages in the educational process. And, we also wish, at some policy level and to some degree, to allocate positions of societal leadership and responsibility, access to which is an important consequence of higher education, on the basis of a valuation of worth and benefit to the soceity. These partially overlapping goals impose a difficult and disputatious task on those who would formulate admissions procedures and policies, and especially on

high or much aptitude. Thus, the former will acquire proficiency while the latter will not.

In Carroll's (1963) reformulation, systemization, and extension of the traditional framework for aptitude and its role in school learning (Figure 1), there are at least three other elements which play salient determinative roles in the acquisition of proficiencies via schooling. These are: the individual's understanding of what will be required of him when he undertakes the learning task, his persistence or perseverence in pursuing and fulfilling these requirements, and the amount of time he is allowed to fulfill them.

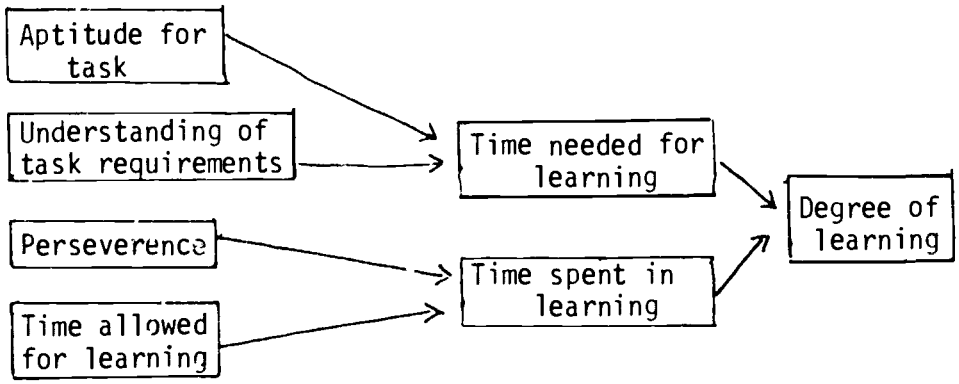

Figure 1. Key Elements in Carroll's Model of School Learning (Adapted from Harnischfeger and Wiley, 1978)

Two points about Carroll's articulation are directly relevant here. First, the "aptitude" which is central to the model is specific to the educational task at hand. Carroll postulates more "basic" aptitudes, but these are "mixed" in varying proportions depending on the task and are "diluted" in their impact on the directly relevant

capacities by specific prior learnings. Second, the notion of personal "capacity" is built into the aptitude concept directly through the "mixing" notion and indirectly by separately accounting the roles of motivation (via perseverence), opportunity (via time allowed for learning), and instructional quality (via perseverence and understanding of task).

It is the "capacity" notion which relates most directly to those of the report's objections which focus on the historical relations of Brigham and the SAT to the mental testing movement. How "basic" are these capacities for learning? This depends strongly on the kind of training, the scope of the learning tasks, and the perspective one adopts about the psychological and social processes central to learnings. Carroll's conceptual framework can be used to encompass a variety of educational experiences and a considerable variation in viewpoint.

And these viewpoints do vary significantly. Thus, Jensen (e.g., 1973) sees "intelligence" as being constituted of a small number of basic learning aptitudes (perhaps only two) which are directly identified with learning abilities. Carroll himself adopts a more differentiated multifactor view with explicit emphasis on specific prior learnings as a component of task-specific aptitude. Cronbach and Snow (1977) eschew any emphasis on small numbers of more "basic" characteristics taking their impetus from the tasks themselves, agnostic to the generality of the concept. For them, even a single psychological task

is too complex to accommodate a single role for an "aptitude" and they emphasize how such tasks may themselves be arranged to accommodate diverse patterns of psychological characteristics. Bloom (1976) implicitly throws out the whole concept of aptitude, claiming primacy of motivation and instructional history--both in home and school. And Cole (e.g., Cole, et al., 1971) explicitly rejects the notions of intelligence and aptitude, treating all learnings as products of specific socializing experiences and implicitly viewing "capacities" as equal, with the social and cultural context of prior learnings determining their content and therefore their relevance as prerequisites for new tasks.

As the educational task becomes more variegated, consumes more time, and occurs later in the educational process, these perspectives become more difficult to apply in organizing a way of thinking about "capacity to acquire proficiency." Is ability to use a library to find appropriate reference materials an aptitude for college education? Surely it satisfies the traditional definition. In Carroll's terms, it is presumably a prior learning which forms part of the task-specific aptitude. If so, then a "SAT" used for college admission might contain measures of prerequisite achievements as well as attempting the measurement of psychological characteristics thought to contribute more "fundamentally" to learning rate.

How does the Educational Testing Service actually view the "aptitudes" which its tests attempt to measure? E.g., what are the "developed abilities" which are designed into the Scholastic Aptitude Test? How much of the test is devoted to aptitude in the sense of forecasting those who will "do well"--whether because of college prerequisite or curricularly-relevant prior achievement or because of more "fundamental" abilities? In particular, how does the actual content of the SAT relate to the report's charges of a "claimed scale of uniform, universal mental merit" in the context of psychologists' core conception of "aptitude"?

In the recently released version of the SAT (Educational Testing Service, 1978), the item allocations to content are:

Verbal	85	
Antonyms		25
Analogies		20
Sentence Completion		15
Reading Passages		25
Mathematical	60	
Algebraic, Arithmetic and Geometric Problems with Multiple Choice Alternatives		40
Quantitative Comparisons		20
Written English	50	
Usage		35
Sentence Correction		15
Total	195	

As the verbal, mathematical, and writing scores are separately reported, the item allocations across these gross categories relate primarily to accuracy of the assessments rather than the "content balance" of a single score. Some inadequacies are obvious, e.g., the test of written English omits all but grammar and usage, the most basic of writing skill areas. And this is an interesting example.

The written English test itself is a new addition to the SAT. It was added to redress criticism that an important area of skills required of college students had been omitted, but was created as a placement rather than a selection test. However, it reflects the inadequacy of existing test technology to provide cost-effective assessments of some skills. The grammatical knowledge and skill of the student is much easier to measure via the multiple choice format than more complex and holistic writing abilities. ETS has concluded that it could not accurately measure these abilities using the cost-efficient multiple-choice technology underlying their existing admissions testing program.

The mathematical section of the SAT is, at its root, based on a conception of the high school mathematics curriculum and its relation to beginning college mathematics and science. It primarily covers arithmetic, algebra, and geometry, spanning many of the common elements of all college preparatory curricula. In terms of content balance, it would have to be judged in curricular terms, either via a specification of desirable or actual college requirements or of high school preparation.

The verbal section of the SAT presents a very different picture. It does not relate directly to any single part of the high school or college curriculum. The passage reading section is a direct extension of the traditional reading comprehension test. It has been extended to more complex and advanced content and balanced over substantive curricular areas (Narrative, Science, Argumentation, Humanities, Social Science). It emphasizes inference and evaluation to a greater degree than tests given to assess achievement in earlier schooling. The analogies, antonyms, and sentence completion sections are all focussed on vacabulary knowledge but attempt to assess more complex semantic content and relations than do school-based achievement test batteries. Also, as this content becomes more complex, the items tend increasingly to resemble those on some tests of "conceptual ability" rather than traditional vocabulary items.

Surely all of mastered content has been previously learned by the students who take the examination. None of it has fallen like manna from heaven. But this fact does not, by itself, invalidate claims that "basic" abilities are assessed. For psychologists studying human abilities, the fundamental task has always been to address the issue of the "basic" and the "general" using evidence deduced from behavior on specific tasks resulting in current performance.

More critically, recent research by Messick (1980)--conducted within and sponsored by ETS--indicates that even short-term instruction can have considerable impact on SAT scores. How much of these gains occur on items which relate to specific courses in the high school curriculum and how much to items--like the verbal analogies--which have traditionally been thought of as indices of more "basic" processes, has not been revealed. Clues are available, however, as Messick's analyses exhibit much greater impacts of particular amounts of instruction on the mathematical section, which has much more course-related content, than the verbal.

Thus, it would seem that some of the content is more "basic"--in traditional terms--and some is more achievement-like, and thus easier to modify through direct instruction. In more general terms, how much of the score variation reflects components which will save the student effort and time in college because they diagnose already-mastered learnings which others must newly accomplish and how much of that variation reflects more general abilities which speed new learnings is surely relevant to the selection policies implemented with the instrument, but about which no evidence exists currently.

What, then, are the answers to the critical questions: How "fundamental" is this content? How important is it? How relevant is it to college admission?

How much of the content of a Scholastic Aptitude Test is "fundamental" depends on one's views concerning aptitude as a concept as well as empirical evidence concerning the modifiability of the scores and their components. Currently, there is wide variation in perspective among members of the scholarly community and only the beginnings of a reasonable empirical analysis.

The manifest content is obviously important, but is--also obviously--incomplete. And its relevance is surely subject to disagreement and dispute. Much of the incompleteness and some of the disagreement results from inadequacies in the testing technology, but much of both also results from lack of open debate on the importance of various skills and capacities as we screen and select individuals for those opportunities which lead to positions of societal leadership. These concerns are not only those of the colleges to whom we admit our students, they are also the legitimate focus of the citizenry as it engages in political debate over access to leadership in the society. In my view, we can live without completeness--we must and we do--but we cannot live without reflection and debate over the central issues.

One aspect of the report which has received wide attention relates at its core to these issues of test content. The relation between the social background of individuals and their SAT scores is used in the report to support a charge which has nettled ETS (Educational Testing Service, 1980a) and worried many of those who are members of the psycho-

metric community. The charge is that youth from working class and minority backgrounds are illegitimately excluded from life success and societal influence by the tests because of biased selection of skills that are tested and invalidity in methods used to test them. And, additionally, to the extent to which the tests do reflect differences in academically-relevant skills, the selection procedures within which the tests are embedded are said to reinforce the very reasons for existing group differentials: racial discrimination and economically-based restrictions in educational opportunity.

The argument made by ETS to rebut these charges has several elements.

1 The relation of social background to the test scores is smaller than the report asserts it is;

2. the actual procedures used for college admissions decisions utilize more information than the SAT and incorporate financial aid provisions which increase access to colleges of those from families with low income;

3. the introduction of admissions tests such as the SAT, has actually reduced the relation between social background and college attendance; and

4. the primary reason for test-score differentials among individuals with different social backgrounds is that those individuals actually differ in the abilities which the tests measure.

The dispute over the precise magnitude of the social class-test score relation is irrelevant to the basic issue. The relation is substantial in magnitude and the well-known black/white differential in tested ability is quite enormous. The fact that the scores are embedded in a complex and institutionally varying admissions process which incorporates other criteria as well is equally irrelevant. To the extent that test scores are used in the process--which no one disputes--the role they play is a legitimate focus of policy discussion.

The contention that the advent of national admissions testing improved access to post-secondary education for those from economically disadvantaged backgrounds is extremely relevant to the issue, however. Unfortunately, it is precisely this issue on which the two sides disagree and neither the report nor ETS present any evidence to justify their contentions. Each side seems to assume that either (a) it is obvious that testing cuts off access of those who legitimately deserve admission, or (b) that testing has improved such access over the admissions system previously in force. Logically, both assertions could be true as they implicitly contrast the current system with different alternatives: an unspecified "old system" and an equally unspecified new alternative. Perhaps more importantly, however, what each side assumes is obvious may not be true, even if the alternatives were well specified. The conventional wisdom of the testing community, that testing has improved access to college of the socially and econ-

mically disadvantaged, is not supported by any scientifically respectable evidence. The fact that opportunities to attend college have expanded dramatically since the introduction of national admissions testing by no means implies that testing is the cause. The contrasting degree of "rigor" on which ETS has insisted in accepting evidence of the effects of test coaching vs. their criteria for asserting the egalitarian impact of the national admissions testings, leaves one open-mouthed. And, on the part of the report, the case studies and other assertions of harm could surely be matched by similar case studies and assertions of benefit, with equal credibility.

The most important issue, once technically resolvable questions of item and test validity are stripped away, is that of the meaningfulness and significance of real differences in the abilities which the tests are intended to measure. The report charges that these abilities are simultaneously trivial, narrowly defined, and culturally selective. Presumably, this is taken by the report's author to imply that if the abilities were (a) selected to be more central to and representative of the whole range of skills required by meaningful college curricula, and (b) assessed in terms more familiar to those from non-majority and economically-disadvantaged backgrounds, much of the differential would disappear. From my perspective, the great part of this issue is not technical. It is a political issue which focusses on which skills we want to be relevant for college.

We can and have changed the nature of the college-going experience substantially as we increased access throughout this century. In fact, changes in the nature of college and in the character of the selection criteria themselves have resulted in historic changes in the selection process. Thus, some of the basis for the introduction of the modern SAT can probably be traced to (then) partially new criteria which we, as a society, had evolved for access to further education. What the report is actually calling for is a ratification and extension of some of the changes in admissions criteria initiated in the 1960's.

The era of open admissions and alternative criteria constituted a partial rejection of contemporary academic criteria as a basis for college admission. The SAT surely does represent a central distillation of those core academic criteria that can be measured through multiple-choice items. If one disagrees, as Nairn and others do, with those particular core criteria--either because one favors multiple, rather than academically unitary, criteria or because one wants to establish new criteria which will give groups, which currently have low admissability, higher priority--then it is logical to attack the existing system and its embedded criteria. Such disagreements cannot be resolved on technical grounds.

For me, the central issues concerning the role and importance of the

skills assessed do not lie with the content specifications of the current SAT or similar admissions tests. Rather, they lie with the openness and thoughtfulness with which the content decisions are made and a realization that these decisions are not merely technical, they are educational and political ones which reflect our conception of higher education and the role it plays in the society. There will always be dissension over the test content relevant for college admission. And, it will take considerable time for a scientific consensus to emerge about the meaning, centrality, and modifiability of "aptitude." From my perspective, a more open process would improve the societal worth of admission decisions and stimulate relevant research and scientific consensus.

Currently, what we need is an open procedure by which decisions are made which

(1) allocate test components between the more abstract, less specific "aptitude" content and the more curricularly-relevant, specific prior learnings, i.e., achievements;

(2) encourage broad-based psychometric and educational critiques of the kinds of skills and modes of measurement of those classified as "basic"; and

(3) perform a careful sifting and screening of the current secondary curriculum and the (perhaps) consensually appropriate prerequisites for adequate college learnings, in order to set clear specifications of common achievement content.

Additionally, we need clear empirical information on the admissions consequences of the resulting tests for high school students with manifestly meritorious accomplishment and for individuals in groups for which the society has made a commitment to enhance access to positions of societal leadership. Only thus can we, as a society, balance our priorities concerning enhancement of access and social mobility, encouraging diligence and hard work, and keep our educational investments to a satisfactory level of return.

The tests: their validity

Most recent psychometric work on validity-related matters has focussed on the use of tests for selection decisions.[1] This work has been

[1]The report has severely criticized ETS for "falsely" asserting the high predictive validity of the SAT and other instruments. ETS has responded to these criticisms (Educational Testing Service, 1980b) in a lengthy (27 page) printed report. The content of the criticism has to do with substitution of an alternate index of strength of statistical relation for the currently popular one--multiple correlation---used by ETS and the testing field generally to link test scores and high school grades to first year college grades. The most fascinating aspect of the discussion is the widely known fact that both the currently popular index and the report's transformation of it--which reduces its apparent magnitude--are extremely faulty indices of accuracy of prediction. The more adequate base, focussing on misclassifications and incorrect admissions decisions and valuation of their consequences, has been available for some time (Cronbach and Gleser, 1957). It is no wonder that a non-statistician like Nairn is confused if ETS persists in using misleading and arbitrary indices.

strongly stimulated by legal concerns about the fairness of selection procedures; primarily those used in the employment process. The focus of this research has not been on the nature of the tests themselves or the measurements deriving from them, but on the social selection procedures that incorporate these tests. Thus, the implications of the work for changes in the process relate only to the ways in which the scores of individuals with different non-test characteristics are incorporated into the criteria for selection, not to such issues as item content, item format, method of scoring, etc. This research has, following the earlier lead of Cronbach and Gleser (1957), turned the question of test validity into a series of questions relating to the value of the scores, as given, for a set of alternative selection procedures incorporating them. From this perspective, the validity of the test becomes the validities of alternative uses and, epistomologically, the label attached to the test--e.g., reading comprehension, scholastic aptitude, mathematics concepts--plays no formal role in determining its validity. Thus, any test, regardless of original intent, could be used for admission and its validity would depend only on its empirical relations to predictive criteria, not on its content or format. In addition, the test--SAT or whatever--has a different validity at, e.g., Harvard than at Northwestern and no questions relating to validity can be raised until the test results are actually used in some fashion. This viewpoint is surely too narrow.

As a general perspective, it fragments the validity concept--as tests are used in different ways--and it forecloses whole classes of questions that relate to item and test format, content selection, scoring and scaling. From my perspective, the new work does not focus on test validity at all. It primarily is a conceptual framework and a set of standards for assessing the social worth of selection procedures incorporating any criteria that are (a) quantitative, and (b) measured with error. Problematically, it focusses primary attention on external criteria and allows those who should be forced to attend to important concerns about the validity of their devices to ignore them.

Inherently, the notion of test validity must rest on two conceptions: (a) that which a test ought to measure and (b) that which a test does measure. It is the discrepancies between the two, somehow defined, that bear on validity. Central theoretical and practical problems for psychometrics are (1) the mode of specification of the ought and (2) the form of expression of the discrepancy. Recent discussions of the validity concept in the psychometric literature (Cronbach, 1971; 1980) have focussed on the word interpretation as the entity which is validated. However, a central interpretation of "interpretation" has, at least since Cronbach and Meehl (1955), centered on the idea of a definition or theoretical conception of what is intended to be measured (i.e., the "construct")--my ought. The problem with the specification of the ought is that, if it occurs at all in the actual world of test

construction--beyond an undefined label--it is formulated in ways that make it difficult to separate valid from invalid components of the measurements.

Cronbach (1971) gives a salient example of a specification of an intent of measurement which highlights this issue of separation:

> Consider further reading comprehension as a trait construct. Suppose that the test presents paragraphs each followed by multiple-choice questions. The paragraphs obviously call for reading and presumably contain the information needed to answer the questions. Can a question about what the test measures arise? It can, if any counterinterpretation may reasonably be advanced. Here are a few counterhypotheses (Vernon, 1962):
>
> 1. The test is given with a time limit. Speed of of reading may contribute appreciably to the score. The publisher claims that the time limit is generous. But is it?
>
> 2. These paragraphs seem abstract and full. Perhaps able readers who have little motivation for academic work make little effort and therefore earn low scores.
>
> 3. The questions seem to call only for recall of facts presented in simple sentences. One wants to measure ability to comprehend at a higher level than word recognition and recall.
>
> 4. Uncommon words appear in the paragraphs. Is the score more a measure of vocabulary than of reading comprehension?
>
> 5. Do the students who earn good scores really demonstrate superior reading or only a superior test-taking strategy? Perhaps the way to earn a good score is to read the questions first and look up the answers in the paragraph.
>
> 6. Perhaps this is a test of information in which a well-informed student can give good responses without reading the paragraphs at all.

> These miscellaneous challenges express fragments of a definition or theoretical conception of reading comprehension that, if stated explicitly, might begin: "The student considered superior in reading comprehension is one who, if acquainted with the words in a paragraph, will be able to derive from the paragraph the same conclusions that other educated readers, previously uninformed on the subject of the paragraph, derive." Just this one sentence separates superior vocabulary, reading speed, information, and other counterhypotheses from the construct, reading comprehension. The construct is not identified with the whole complex practical task of reading, where information and vocabulary surely contribute to success. A distinctive, separate skill is hypothesized. (pp. 463-464)

Cronbach's example implies several things in this context. First, it makes clear that reading comprehension as an intent of measurement is not all things to all persons; it is not speed, vocabulary, test-wiseness, or prior information, regardless of whether these "constructs" contribute to success on the test task itself, other tasks given contemporaneously, or future tasks. If we take this further and realize that such sources of invalidity in the assessment of reading comprehension are (a) themselves valid intents of measurement with other instruments and are (b) irremovable sources of variation in test performance for many "constructs"[2] then two further implications flow

--the problem of test validation, whether focussed on the notion of "interpretation" or not, cannot be shifted

[2]E.g., vocabulary knowledge is a logical prerequisite for appropriate performance on comprehension test tasks. Although variation in performance due to differences in vocabulary can be suppressed by experimental training or selection of common words, it cannot be removed as a source of extraneous (invalid) variation in practical test situations.

entirely to an analysis of test use, and that

--the labeling of the test or the description of what it is intended to measure must be sufficiently precise to allow the separation of components of invalidity from valid variations in performance.

Also, we must note that these sources of invalidity are often positively related to the characteristic that is the intent of measurement. Thus, in the Cronbach example, those who have the skills necessary for "comprehension" of passage content or derivation of correct conclusions, given adequate vocabulary, will also be more likely to have previously acquired that vocabulary knowledge.

Thus, in more general terms, a component of invalidity of a test with a particular intent may well be positively and even substantially related to both a predictive criterion used to "validate" the test's use in a particular selection process, and the "valid" components of the test itself. The critical policy issues then relate to the legitimacy of the various components and the weight that these components receive in the test score incorporated into the selection process. For example, if the reading comprehension section of the SAT actually has substantial components of vocabulary and test-wiseness as well as comprehension, would we--even if all three are positively and independently related to first-year college grades--find this a satisfactory state of affairs? We might not if we felt that the skills making up test-wiseness in a multiple-choice context were not

appropriate criteria for college admission. Some might also be concerned that the vocabulary components of the SAT were, in reality, more heavily weighted than the original intent--as manifested in the numbers of items allocated to the subtests. These issues could become a matter of critical concern if, e.g., Black test takers were more widely separated from whites on test-wiseness or vocabulary components than on reading comprehension per se.

In order to bring this topic back concretely to the criticisms of the report, let us focus on a specific issue: the multiple choice item format. A charge which threads itself through much of the argumentation is that this mode of testing contributes serious distortions to the measurements produced, simultaneously travializing their meaning and adding bias which contributes to group differences in performance. Surely this theme most directly disvalues the test-wiseness components of the measurements. How is the psycnometric community to respond to this charge? Where is the evidence that would clarify the argumentation or allow reasoned judgement? My own view is that there is none. And I believe that there are several fundamental reasons for this.

Almost all psychometric research, until recently, has been focussed on issues of error and reliability rather than on bias and validity. The theoretical framework for the analysis of measurement errors has become conceptually sophisticated, elaborate and full of concrete

detail. It has progressed to the point that primitive correlational indices are no longer scientifically respectable as having clear meaning and where the conceptual and analytic frameworks for test items and responses to them are fully integrated with those for test scores. On the other hand, the conceptual orientations to validity of tests are diffuse, fragmented and fundamentally incomplete. The widely accepted rubric of "construct validity" (Cronbach and Meehl) is abstractive enough so that it gives little or no guidance in the choice of operational procedures or the allocation of investigative resources. The decision-theoretic analysis of selection decisions (Cronbach and Gleser), is not integrated in any fundamental fashion with the construct framework. The recent theoretical work on selection bias builds on the decision frame but again ignores the "construct" issues. In fact, the whole issue of test "bias"--at its heart a phenomenon of differential validity--has never been linked to the core theoretical concepts of validity. Finally, in this area, the frameworks for item assessment have never been fundamentally integrated with those for tests. Thus, "item bias" has no bearing on "test bias" and "content validity," which, at the operational level, seems to mean sampling or selection processes for the items which make up the test, has no relation to test validity, which at the operational level, seems to mean a relation to a single external criterion in the (implicit or explicit) context of a selection decision. The fact these non-overlapping processes can be tenuously linked via the vagaries of "construct validity" does not imply that they could actually be integrated.

Clearly, item response format is an item issue. However, it does not fit with the item-level procedures for validation, because any "population" from which a multiple-choice item could be selected would be a "multiple-choice" population and issues of representativeness, having to do with the multiple-choice format, could not arise. On the other hand, it is difficult to see how a predictive validation in a selection framework could possibly address the issue.

From my perspective, to address the multiple-choice issue as raised in the report, one would e to

(1) specify the characteristic to be measured so that it did not incorporate the skills specifically required for the multiple-choice format,

(2) decompose the measurements deriving from a multiple-choice test into (at least) two components: (a) values representing performance variations independent of (conditioned on) skills relating to knowledge of the choices and (b) values representing skills which are only usable dependent on knowledge of the choices, and

(3) use the decomposition to

(a) break down the total test variation into variations and covariations of these components,

(b) proportionately attribute group differences (e.g., Black vs. White) to one component or the other, and

(c) analyse the predictive relations of the separate components to selection criteria.

These procedures flow immediately from neither the content validity framework nor the selection perspective. Thus, as an issue of psychometric interest they would likely only arise because of strong complaints and polemics from outside the field, not from a "natural" topic of investigation. And this is actually how the new selection models themselves arose. They were not the direct evolute of ongoing research, they came about because of controversy from outside.

The tests: their role in educational selection

The processes by which the work histories of individuals are formed and determined are the most important in life. It is the work of an individual which determines his or her productive contribution to the society at large and which generates the rewards which are instrumental in satisfying the person's needs and wants. The educational system constitutes the formal institution by which society sifts and differentiates, as well as educates, individuals for such work patterns.

The American educational system differentiates individuals, both formally and informally, with respect to their educational experiences. And it does so at all levels of the educational system. At the earliest levels, one finds ability grouping and "individualized" instruction. At the middle levels one finds educational tracking. And at the higher levels one finds differential selection and institutional stratification. Increasingly, in this century, the processes by which these differentiations have been made have, at all levels, incorporated formal procedures for testing the abilities of pupils.

Historically, this increasing reliance on tests has paralleled the great increases in educational levels attained by succeeding age groups of young Americans. These increases have pushed upward those levels at which the selective distinctions ease or bar access to initial occupational roles leading, eventually, to positions of leadership and responsibility. Currently, in my view, the critical decisions controlling access to the higher level occupations and institutional niches are those at the college and professional school entry points.

The major reason for the use of tests within the educational system --other than their presumed validity--is the low cost of current multiple-choice test technology. The fact that individuals mark answer sheets by blackening one or more response areas, so that machines can read, score, analyze, and report results, lowers monetary costs to manageable boundaries for educational institutions and perhaps even students. Thus, the perceived benefit per unit cost is quite high especially for institutions which do not wish to devote many of their resources to these decisions.

When we turn to higher-education admissions processes as one of these critical decision points, the circumstances and alternatives become more fragmented and heterogeneous than at earlier stages. Individuals who were hitherto relatively passive participants in an institutionally-bounded educational decision process, enter an institutionally hetero-

geneous one in which their active participation, via the application process, is required. The interactive institutional selection process then becomes more significant in the eventual access to higher-level positions than does the base decision to enter higher education at all. This continuation decision is currently much more important to eventual occupational and work organization access at intermediate levels of income and influence than is the institutional selection decision.

Most studies of higher education as an aspect of social stratification and mobility have ignored important institutional differences within the post-secondary system. That system can be (perhaps simplistically) summarized--for purposes of undergraduate education--as a three-tier system: elite private colleges and universities, public four-year colleges and universities, two-year public colleges. Within each tier there are also internal strata which differentiate eventual opportunities for further education and access to various strata within the occupational system itself. The criteria for selection into each of the undergraduate tiers are distinct. At the community college level, all that is usually required is a high school diploma. For public four-year colleges and universities, the central and sometimes only criteria are high school grades and test scores. For the elite private colleges and universities, the criteria are more complex and varied. These institutions seek "well rounded" individuals because they are acutely aware that they are preparing

future societal leaders and that intelligence, diligence, and intellectual achievement are not sufficient prerequisites for these roles. These institutions also, however, tolerate considerable variations in profiles because they see a variety of specific skills and behavioral habits as compensating one another over the elite occupational roles for which they are preparing students. They also have been able, in some cases, to recruit significant numbers of minority students because they have evolved commitments to augment the numbers of such individuals in these elite roles. And the greater diffuseness of their admissions criteria aids this process.

In this context, therefore, the defensive claim that admissions tests do not bar individuals from higher education does not address the central issue. The key question concerns the manner in which the tests control access to institutions of specific types or levels within the internally stratified collection of institutions of higher education. If we accept that there are strong linkages of institutional status to the eventual standings of individuals in the economic and political system, then the tests may have powerful impacts on these outcomes.

It is the linkage of standing or "level" of undergraduate institution to "level" of graduate institution and thence to status and influence within an occupation which is the motivation for considering the

institutional standing issue important. What proportion of Harvard undergraduate degree holders continue at Yale Law School and what is their economic standing twenty years hence? Is Harvard "overrepresented" in the U.S. Congress? These linkages are obscured in short-term quantitative studies of the educational determinants of social stratification because of the relative rarity of these events in statistical terms, and because of the difficulties of constructing and linking individual and institutional measurements over the lengths of time necessary for effects to appear. However, they are, even in the absence of social science justification, the basis for major resource allocation decisions within elite colleges and universities.

However, given the reality of that linkage, the issue of how test information controls access to institutions with various academic and social standings becomes central. Thus, books are published giving information on the SAT score levels of those attending particular institutions. And applicant SAT scores are now available to them before the application deadline dates. It is thus not difficult to make credible the potentially powerful contribution of the tests in the stratification process under these circumstances regardless of the actual use of the test data by a particular institution or by institutions in the aggregate. However, we have very little real evidence about overall admissions decision processes, at least as they affect college choice and selection in the aggregate. What

little national evidence we have is focussed primarily on the role of financial aid. And holistic studies, where they exist, are conducted by specific institutions in the process of constructing their "marketing" plans. The total process with its elements of (a) deciding to which institution to apply--based on parental and counselor preference, family financial constraints, high school grades and national test scores, (b) the institutional decision process, and (c) the eventual acceptance decision, has been incompletely studied and the evidential fragments which do exist have never been gathered together.

In my view, it is likely that national admissions tests--particularly the SAT--play central roles in the process of social stratification, especially at the upper levels.

What do I conclude about the report after this long string of analysis and argumentation?

First, that many of the issues raised are socially and technically fundamental. They have to do with the role that tests play in our educational system and in the society at large. The polemic and the rhetoric, at their best, force us to examine and deal with that role from the "outside," rather than the "inside." It forces on the educational and psychometric communities the responsibility of

responding and arguing in broad rather than narrow terms and thus may force consideration of basic issues hitherto obscured by the day-to-day parochial concerns of those communities. For all of us, the nature of tests and their content must become again paramount educational issues. The fact the tests exemplify educational goals has become a truism. We do not sufficiently realize that our tests, especially as they partake of the latter stages of the educational process, must both reflect and form the greater society, promoting and defining its leadership criteria as well as articulating our aspirations for it. Thus, test content cannot be masked as only or solely a technical issue to be decided by the psychometric and educational communities, narrowly defined.

Second, that a great deal of the fragmentation and imprecision in the book's argumentation is due to fragmentation and imprecision in the very core concept of the testing field: validity. Thus, we are currently being forced, not only by this report, but by uncertainty and criticism of the roles which tests play in our educational system and our society, to reconsider what we mean by valid and adequate measurement of human abilities and in fact, what we mean by the abilities themselves. And then there is the dilemma of multiple choice testing technology: can we really evaluate its costs and benefits without having a conception of validity which allows us to ask scientifically answerable questions concerning the magnitude of measurement distortions induced by multiple choice formats?

These are important questions which both we and Nairn are ill prepared to ask precisely, let alone answer.

Finally, a sad evaluation is that the report is so argumentatively and conceptually inadequate, even against its impoverished scientific and policy context, that the valid social functions it could have served will not be fulfilled. The book is a terribly disappointing waste of an enormous and potentially valuable time investment.

References

Bloom, B.S. Human characteristics and school learning. New York: McGraw-Hill, 1976.

Carroll, J.B. A model for school learning. Teachers College Record, 1963, 64, 723-733.

Cole, M., Gay, J., Glick, J.A. & Sharp, D.W. The cultural context of learning and thinking. New York: Basic Books, 1971.

Cronbach, L.J. Test validation. In Thorndike, R.L. Educational measurement, second edition. Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. Validity on parole: how can we go straight? New Directions for Testing and Measurement, 1980, 5, 99-108.

Cronbach, L.J. & Gleser, G.C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1957.

Cronbach, L.J. & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 282-300.

Cronbach, L.J. & Snow, R.E. Aptitudes and instructional methods. New York: Irvington, 1977.

Educational Testing Service. Taking the SAT. Princeton: ETS, 1978.

Educational Testing Service. Test scores and family income. ETS, 1980a.

Educational Testing Service. Test use and validity. ETS, 1980b.

English, H.B. & English, A.C. A comprehensive dictionary of psychological and psychoanalytical terms. New York: McKay, 1958.

Harnischfeger, A. & Wiley, D.E. Conceptual issues in models of school learning. Curriculum Studies, 1978, 10, 215-231.

Jensen, A.R. Educability and group differences. New York: Harper and Row, 1973.

Messick, S. The effectiveness of coaching for the SAT: review and reanalysis of research from the fifties to the FTC. Princeton: ETS, 1980.

Vernon, P.E. The determinants of reading comprehension. Educational and Psychological Measurement. 1962, 22, 269-286.